1. More Visualization Results

Figure 1 illustrates the explanations produced by different methods. Compared to baseline methods, our TokenTM accurately localizes the rationales behind the model's prediction, resulting in more human-understandable interpretations.

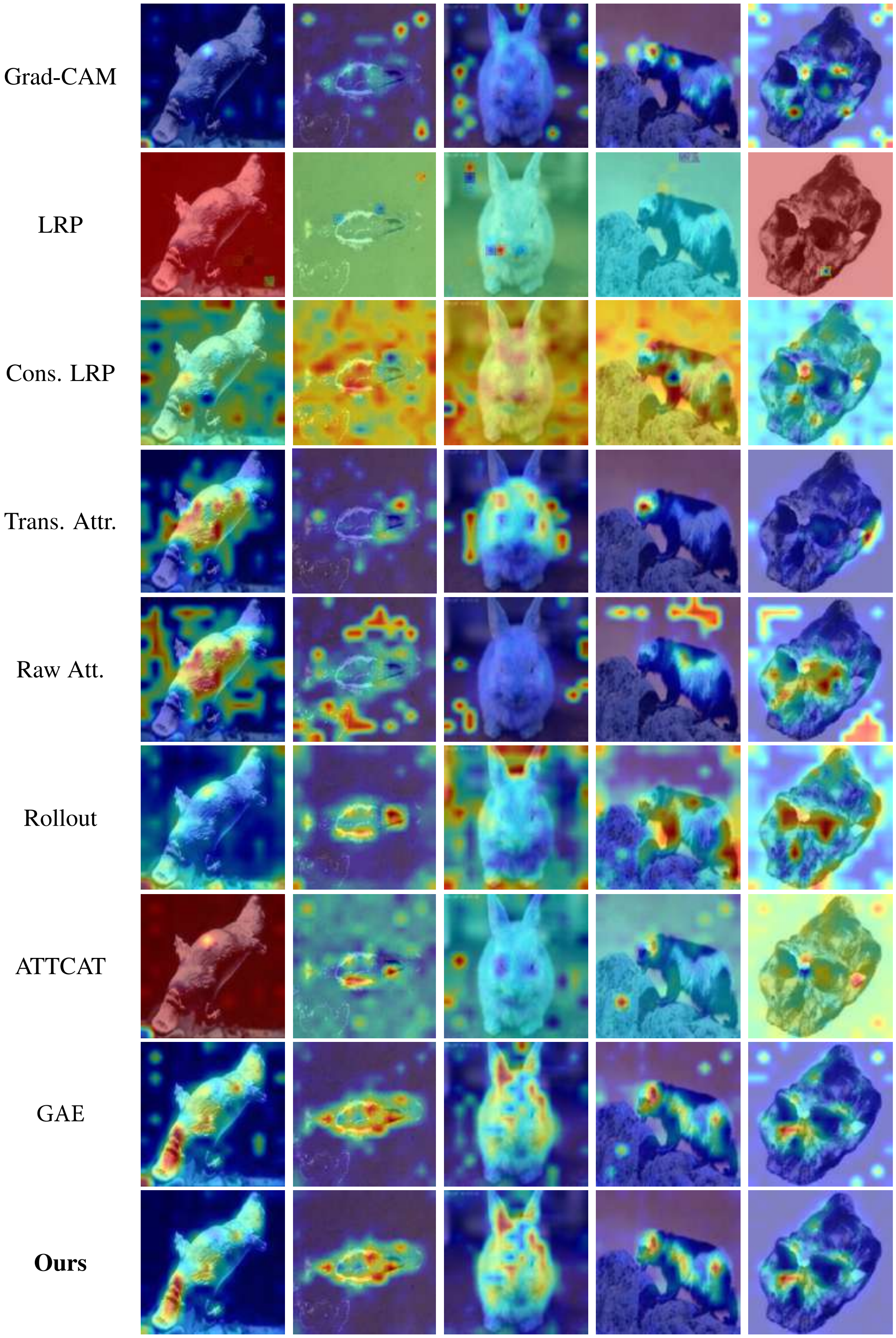

Figure 1. Visualizations of explanation results. Our method produces more object-centric heatmaps.

2. Detail of Experimental Setup

2.1. Datasets

CIFAR-10 and CIFAR-100. CIFAR-10 and CIFAR-100 [10] are two widely used image classification datasets, each containing 60,000 32 × 32 color images. CIFAR-10 has 10 classes, while CIFAR-100 has a more challenging setting with 100 classes. Both datasets are split into 50,000 training and 10,000 testing images. In this paper, we evaluate explanation methods on the testing sets.

ImageNet. ImageNet [14] is a large-scale benchmark for image classification. In this work, we evaluate explanation methods on the validation set, which comprises 50,000 high-resolution images across 1,000 distinct classes. Each class contains roughly the same number of images, ensuring a balanced benchmark.

ImageNet-Segmentation. ImageNet-Segmentation [8] is a subset of ImageNet with segmentation annotations, containing 4,276 images from 445 categories.

2.2. Implementation of Baseline Methods

2.2.1 Gradient-based Methods

Grad-CAM. Grad-CAM [15] considers the last attention map and utilizes the row corresponding to the $[CLS]$ token, which is then mapped onto the 2D image space. Different from Raw Attention, Grad-CAM performs multi-head integration using the gradient. We implement this method on Vision Transformers following previous works [5, 6].

2.2.2 Attribution-based Methods

LRP. LRP [4] starts from the model's output and propagates relevance scores backward up to the input image. The propagation adheres to a set of rules defined by the Deep Taylor Decomposition theory [11].

Conservative LRP. Conservative LRP [2] introduces specialized LRP rules for attention heads and layer norms in Transformer models. This is designed to implement conservation, a desirable property of attribution-based techniques.

Transformer Attribution. Transformer Attribution [6] is an attribution-based method that is specifically designed for Transformer models. It first computes relevance scores via modified LRP and then integrates these scores with attention maps to produce an explanation.

2.2.3 Attention-based Methods

Raw Attention. Raw Attention [9] extracts the multi-head attention map from the last layer of the model and reshapes the row corresponding to the $[CLS]$ token into the 2D image space. The explanation result is further obtained by averaging across different heads.

Rollout. Rollout [1] interprets the information flow within Transformers from the perspective of Directed Acyclic Graphs (DAGs). It traces and accumulates the attention weights across various layers using a linear combination strategy.

ATTCAT. ATTCAT [13] is a Transformer explanation technique using attentive class activation tokens. It employs a combination of encoded features, their associated gradients, and their attention weights to produce confident explanations.

GAE. GAE [5] is a general interpretation framework applicable to diverse Transformer architectures. It aggregates attention maps with corresponding gradients to generate class-specific explanations.

2.3. Evaluation Metrics

Area Under the Curve (AUC) ↓. This metric calculates the Area Under the Curve (AUC) corresponding to the model's performance as different proportions of input pixels are perturbed [3]. To elaborate, we first generate new data by gradually removing pixels in increments of 5% (from 0% to 100%) based on their explanation weights. The model's accuracy is then assessed on these perturbed data, resulting in a sequence of accuracy measurements. The AUC is subsequently computed using this sequence.

Area Over the Perturbation Curve (AOPC) ↑. AOPC [7, 12] measures the changes in output probabilities *w.r.t.* the predicted label after perturbations:

$$\text{AOPC} = \frac{1}{|\mathbb{K}|} \sum_{k \in \mathbb{K}} (\hat{p}(y|\mathbf{x}) - \hat{p}(y|\mathbf{x_k})), \tag{1}$$

where $\mathbb{K} = \{0, 5, ..., 95, 100\}$ is a set of perturbation levels, $\hat{p}(y|\mathbf{x})$ estimates the probability for the predicted class given a sample $\mathbf{x}$, and $\mathbf{x_k}$ is the perturbed version of $\mathbf{x}$, from which the top $k\%$ pixels ranked by explanation weights are removed.

Log-odds score (LOdds) ↓. LOdds [13, 16] averages the difference between the negative logarithmic probabilities on the predicted label before and after masking $k\%$ top-scored pixels over the perturbation set $\mathbb{K}$:

$$\text{LOdds} = -\frac{1}{|\mathbb{K}|}\sum_{k\in\mathbb{K}} \log \frac{\hat{p}(y|\mathbf{x})}{\hat{p}(y|\mathbf{x_k})}. \tag{2}$$

The notations are the same as in Eq. (1).

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020. 2

[2] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *ICML*, 2022. 2

[3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *EMNLP*, 2020. 2

[4] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *ICANN*, 2016. 2

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 2

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 2

[7] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *ACL*, 2020. 2

[8] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *IJCV*, 110: 328–348, 2014. 2

[9] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *NAACL*, 2019. 2

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[11] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *PR*, 65:211–222, 2017. 2

[12] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL*, 2018. 2

[13] Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining transformers via attentive class activation tokens. In *NeurIPS*, 2022. 2, 3

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 2

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 2

[16] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017. 3